Lab 2: Decision Trees

20.03.2017

Laboratories are aimed at practicing the induction of decision tree models.

1. Induce a decision tree from given set of learning examples (perform manual calculations supported by some tools).

Hangover	Exam	Weekend	Party
no	easy	no	yes
no	hard	no	no
no	no	no	yes
no	no	yes	yes
yes	easy	no	no
yes	hard	no	no
yes	no	no	no
yes	no	yes	no

2. Split the following attribute of age using a minimum entropy approach:

Age	38	53	21	84	42	55	17	5	88	61	10	7	13
Class	T	T	N	N	T	N	T	N	N	T	N	T	T

3. You can use an original version of C4.5 with Windows interface (done by P.Cichy at PP) – download it from the following page http://www.cs.put.poznan.pl/mkomosinski/umsn/c45.zip

   - Unpack *C4_5.exe*; Study help files - *c4_5.hlp* i *c45_help.txt*.
   - Use the examples available in zip - golf.nam, golf.dat / analyze their context using text viewer.
   - Induce a tree from golf data - analyze both pruned and unpruned trees (do you see any differences?); Examine contents of confusion matrix.
   - switch between info-gain and gain-ratio splitting criterion - do you see any difference?
   - For the same data (golf) - transform the tree into rules - compare results to rules.

4. Use this C45 implementation to induce trees from more complex data - e.g. choose crx data

   - check usefulness of switching between info-gain and gain-ratio splitting criterion
   - induce a binary tree (with grouping values)
   - Analyse differences between these two ways of dealing with highly branching attributes

5. Study the role of pruning for a noisy data monk2 data.

   - Induce the tree from this data - analyze results (structure of both pruned and unpruned trees)
   - Carry out cross-validation procedure. Study the obtained accuracies and confusion matrix.
   - Decide which tree (pruned or unpruned) is a better classifier.
   - Examine other degrees of pruning - find the parameter called (Pruning confidence level). Change its standard value to smaller and larger (e.g. you can make a systematic sensitivity analysis of changing it from 0.05 till 0.5 - with a step 0.05 - and using 10-fold cross validation register average classification accuracy. Make a report - size of the tree (number of nodes and leaves) vs. its classification accuracy. Find the best compromise between the size and predictive ability of the tree.

6. ====================================

   Study slides with the Introduction to WEKA to learn how to use it.

7. Download the attached data set - contact lenses - as an example to play with WEKA

   - Read the file and see the characteristics of its attributes
   - Induce a tree with J4.8 classifier
   - Study obtained results (tree structure and its classification evaluation)

8. Create an arff file from the set of learning examples given in Exercise 1 (party table).

9. Induce a decision tree in WEKA software using J48 classifier (unprunned version). Compare the result with the one from Exercise 1. Compare the prunned one with the unprunned version. Is there any difference?

10. Go to the UC Irvine Machine Learning Repository: UCI ML Repository.

11. Choose 2 different data sets concerning topics that are interesting for you.

12. Prepare arff files with chosen data.

13. Induce a decision tree in WEKA software using J48 classifier (unprunned version and prunned version). Test different parameters for prunning. Compare the obtained results. Is there any difference? How pruning affects the model with respect to: the size of the tree, the training error and the testing error (estimated by 10-fold CV).